**Санкт-Петербургский государственный университет Факультет прикладной математики и процессов управления**

**Лабораторная работа №3**

**Обезличивание дата-сета**

**Вариант - 4**

|  |  |  |
| --- | --- | --- |
| Студент гр. Б15-ПУ |  | Шайдуров В.Д. |
| Преподаватель |  | Щеголева Н.Л. |

Санкт-Петербург

2022 г.

**Оглавление**

[**1. Цель работы 3**](#_heading=h.gjdgxs)

[**2. Задача 3**](#_heading=h.30j0zll)

[**3. Описание программы 3**](#_heading=h.1fob9te)

[**3.1. Теоретические сведения 3**](#_heading=h.3znysh7)

[**3.2. Описание функций 4**](#_heading=h.2et92p0)

[**3.3. Общий ход программы 5**](#_heading=h.tyjcwt)

[**4. Оценка полезности данных 7**](#_heading=h.3dy6vkm)

[**5. Рекомендации программиста 8**](#_heading=h.1t3h5sf)

[**6. Вывод 9**](#_heading=h.4d34og8)

[**7. Ссылки: 1**](#_heading=h.2s8eyo1)**0**

# **Цель работы**

Обезличить исходный дата-сет и оценить его информативность.

# **Задача**

На примере дата-сета, созданного в лабораторной работе №2, обезличить данные применяя разные способы анонимизации и вычислить его параметр k-anonymity.

# **Описание программы**

# **Теоретические сведения**

Чтобы оценить защищённость обезличенных данных, был придуман термин k-anonymity, который отвечает за число повторяющихся строк. Если k-anonymity = 1, значит в обезличенном наборе существуют уникальные строки, по которым можно вычислить определённого атрибута, если k-anonymity = 2, тогда во всём дата-сета для каждой строки найдётся минимум одна идентичная строка.

В ходе выполнения данной работы были применены следующие

способы анонимизации данных:

1. **Агрегация** (Обобщение) – техника защиты основана на построении обобщений по одному или нескольким атрибутам (групповые операции). Представляет из себя построение обобщенных данных по выделенным атрибутам: “Лечащий врач”, “Симптомы и болезни” и “Анализы”. Так как последние два столбца в большинстве случаев могут быть уникальными для каждого посетителя, составляется таблица statistics.csv, в которой каждому врачу сопоставляется наиболее встречающаяся симптоматика и назначаемые анализы.
2. **Маскеризация** (Обобщение) – техника защиты, которая скрывает часть индивидуальной информации оставляя наиболее общую для каждого атрибута. Была применена к столбцу “Оплачено с”, чтобы скрыть индивидуальный номер банковской карты клиента.
3. **Удаление атрибутов**  (Подавление) – удаления прямых идентификаторов, а также удаления избыточных квази-идентификаторов, таких как: “Паспортные данные”, “Снилс”, “Симптомы и болезни”, “Анализы”, “Время получения анализов”.
4. **Локальное подавление** (Подавление) – удаление уникальных строк, для достижения нужного параметра k-anonymity. Применяется для составления таблицы anonymity\_data.csv.
5. **Локальное обобщение** (Обобщение) – в рамках данной техники, предполагается уменьшение специфичности атрибута за счет подмены точного значения атрибута его общим значением. Используется для столбца “Дата регистрации”, заменяя точную дату на время года и для столбца “Чек в ₽”, где число округляется до n разряда.
6. **Синтетические данные** (Псевдонимизация) – создает поддельные данные. Такой метод был использован для столбца “ФИО”, для которого реальные значения заменяются на синтетические псевдонимы. Хоть данный метод и несёт смысловой нагрузки дата-сету, он призван помешать злоумышленнику в попытке деанонимизировать дата-сет.
7. **Перемешивание** (Рандомизация) – техника направлена на перемешивание аналогичных записей при сохранении статической значимости всего набора. Применяется к столбцам “Дата регистрации”, “Лечащий врач”, “Чек в ₽”, “Оплачено с” с использованием уникального ключа задаваемым пользователем, необходимым для восстановления порядка в случае деанонимизации.

Всего существует пять основных техник обезличивания, таких как: обобщение, псевдонимизация, рандомизация, подавление и шифрование. Как видно, мною был полностью не задействован метод шифрования, это было сделано специально, так-как оно может применяться для защиты прямых идентификаторов и не вполне является методом обезличивания.

# **Описание функций**

В таблице 3.2.1 представлены использованные библиотеки.

Таблица 3.2.1

|  |  |
| --- | --- |
| Название библиотеки | Описание |
| faker | Помогает создать синтетические данные |
| pycanon | Производит подсчёт k-anonymity |
| sklearn.utils | Вызывает метод shuffle чтобы перемешивать данные |

В таблице 3.2.2 представлено описание функций.

Таблица 3.2.2

|  |  |  |
| --- | --- | --- |
| Имя функции | Входные  Данные | Описание функции |
| med\_aggregation() | - | Локальное подавление над частью исходного дата-сета, из которого мы получаем информацию о каждом докторе и их родом деятельности, такую как: семантика наиболее встречающихся симптомов и наиболее назначаемых анализов. |
| hide\_date() | t – дата посещения  врача | Применяет локальное обобщение для даты посещения врача, заменяется точную дату на время года. |

# **Общий ход программы**

Для дата-сета пользователь задаёт необходимое число уникальных строк (параметр k – anonymity), нужное чтобы регулировать соотношение полезной информации к защищенности персональных данных, которые частично возможно восстановить при помощи квази-идентификаторов.

Вызывается функция med\_aggregation(), чтобы составить новый обобщённый дата-сет statistics.csv, который может понадобится для будущей аналитики (см таблицу 3.2.2).

Далее создаём вторую таблицу данных, в которой удаляются все уникальные идентификаторы, такие как паспорт, снилс, симптомы больного и анализы. Также вместе с уникальными идентификаторами удаляется время получения анализов, которое хоть и является квази-идентификатором, не несёт в себе никакой полезной нагрузки, а только способно сильнее подорвать уникальность данных. Так-как промежуток времени посещения больницы до получения анализов сильно ограничен, а именно двумя днями, и может приходится на смену времён года.

Потом применяются разные способы обезличивания над квази-идентификаторами, такими как: дата регистрации, лечащий врач, чек и метод с помощью которого была произведена оплата (см пункт 3.2).

В то же время фио посетителя заменяется синтетическим псевдонимом.

Происходит подсчёт параметров k – anonymity и удаляются наиболее уникальные строки, параметр которых ниже порогового значения введённого пользователем. Таким образом создаётся новая таблица anonymity\_data.csv.

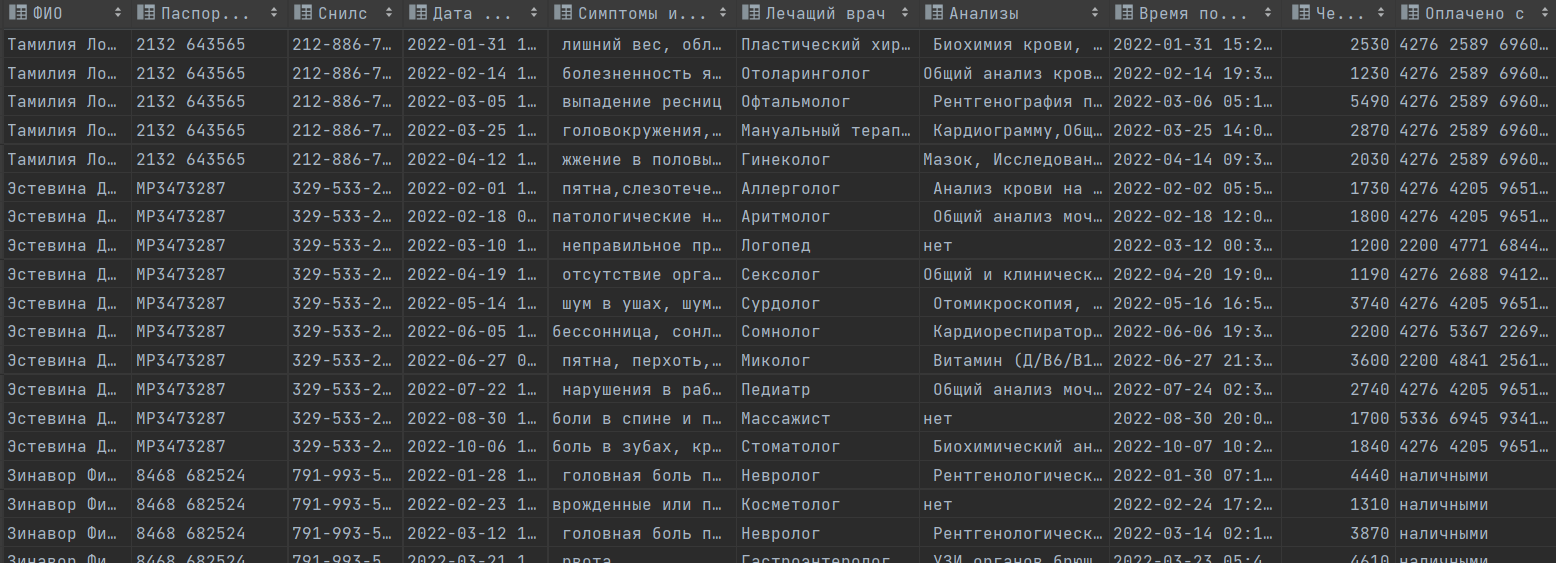


Рисунок 3.3.1. Пример исходной таблицы data.csv

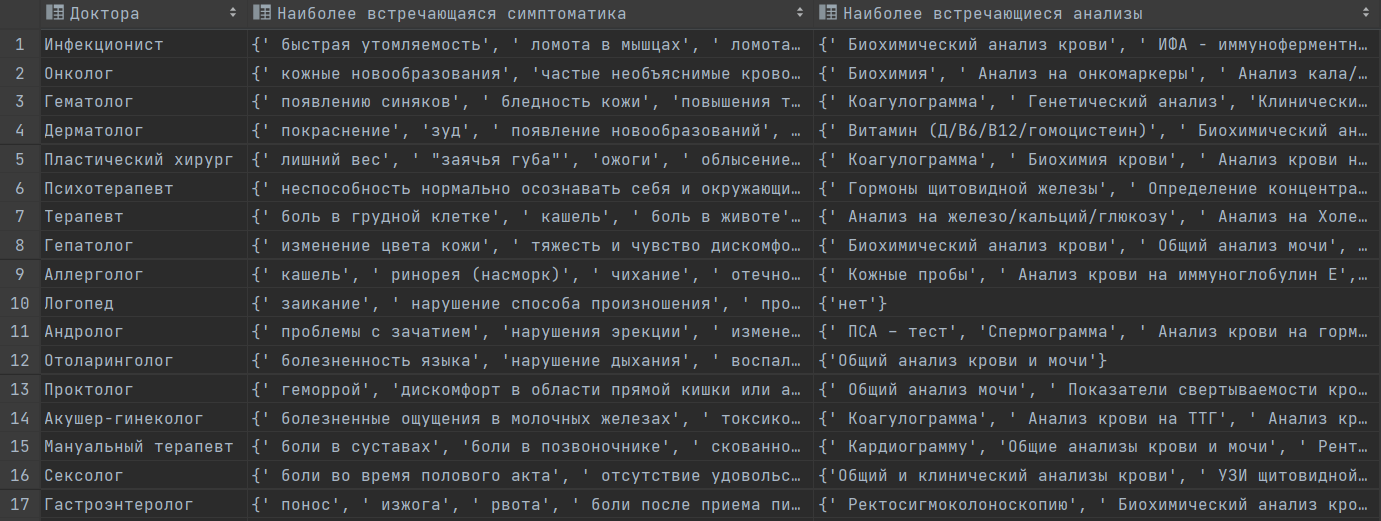


Рисунок 3.3.2. Пример таблицы statistics.csv

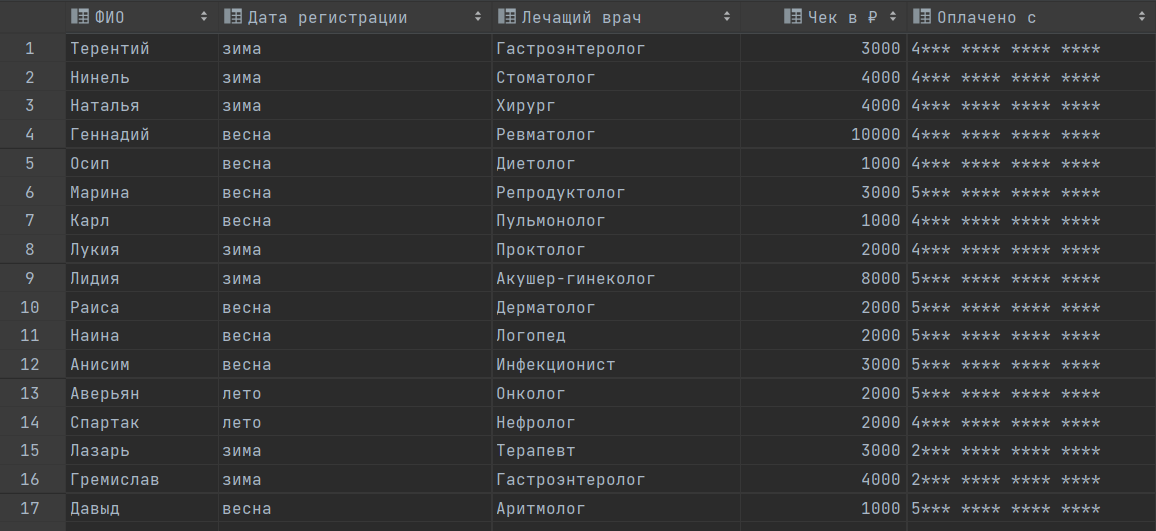


Рисунок 3.3.3. Пример таблицы anonymity\_data.csv

# **Оценка полезности данных**

Программе можно задавать параметр k – anonymity, ключи перемешивания для каждого столбца, округление до нужного разряда, и количество доступных первых цифр для банковской карты. Естественно все эти параметры влияют на формирование и информативность дата-сета, поэтому в конце выводится процент удалённых строк не удовлетворяющих необходимому параметру k – anonymity.Таким образом конечное качество таблицы пользователь определяет самостоятельно.

Если говорить о информации, которую мы теряем, то это: гендерная составляющая, точная информация которая могла бы соответствовать конкретным датам (день, месяц, год), информация о расовой принадлежности гражданина (Россия, Казахстан, Белоруссия), соответствие определённой группы посетителей к имеющимся у них заболеваниям и сданным анализам.

Необходимо заметить что добавление подобной утончающей информации неизбежно будет приводить к увеличению уникальных строк и будет труднее добиться необходимой анонимности без сильной потери атрибутов.

# **Рекомендации программиста**

Для запуска программы необходима 64-битная операционная система Windows и рекомендованная версия python 3.7. Для работы с кодом необходима IDE PyCharm или другая любая среда разработки для python.

Не забудьте импортировать библиотеки см. таблицу 3.2.1.

# **Вывод**

В ходе выполнения данной работы на примере дата-сета из предыдущей л.р.№2 были исследованы методы обезличивания данных, a также произведён анализ полезности получившегося обезличенного набора данных.

# **Ссылки:**

1. Ознакомиться с программной реализацией можно по ссылке:  [git\_hub](https://github.com/Mooncake911/Generation-of-medical-data.git)

2. Подробнее про методы обезличивания данных можно прочитать здесь: <https://sdcpractice.readthedocs.io/en/latest/anon_methods.html>